Gene×Environment Interaction from Case-Control and Case-Case Approaches

Yan Bai, Lynn R. Goldin, and Alisa M. Goldstein

Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland

Using the Genetic Analysis Workshop 12 data, we applied case-control and case-case approaches to study the effects of a major gene and its interaction with sex on the disease liability. Although no joint additive effect was simulated, the case-case approach detected a small but significant multiplicative interaction effect, which could not be explained by the effect of random error. Given that analyses of "real" data will not be made with the knowledge of the true effects a priori, this study showed that the measure of gene×environment interaction is critical and the definition of interaction should be explicit. © 2001Wiley-Liss, Inc.

Key words: case-case, case-control, gene×environment interaction

INTRODUCTION

Traditional epidemiological designs provide simple and effective ways to look at the association between a disease of interest and both genetic and environmental risk factors [Susser et al., 1989; Khoury et al., 1994]. In recent years, new approaches such as the case-case design have been proposed to study gene×environment interaction [Piegorsch et al., 1994; Khoury and Flanders, 1996]. The case-case design uses only subjects with the disorder of interest. It has substantially more power to detect interaction than the case-control approach. Under the assumption of independence between the genetic and environmental factors, this design offers better precision for estimating interactions than those based on both case and control subjects. However, the case-case design can only measure interaction as a departure from multiplicative effects. That is, it cannot detect gene×environment interaction models with departures from an additive scale.

In the Genetic Analysis Workshop (GAW) 12 data set, both the disease status and quantitative disease-related traits as well as other environmental variables were provided for each subject in the population. With knowledge of the simulation results, the purpose of this study was to investigate the effects of the environmental factors E1 (continuous)

Address reprint requests to Dr. Yan Bai, Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd., Bethesda, MD 20892.

and E2 (dichotomized), the genetic factor (gene 1), and its interaction with sex on the disease liability. Sex was chosen because a gene 1×sex interaction effect was illustrated in a combined segregation-linkage analysis at the workshop [Martinez et al., 2001]. Our goal was to see whether traditional epidemiologic methods would also identify the effects of interest in a family-based study population. We illustrated this approach by using subjects from the population isolate.

METHODS

Sampling Procedures

We used subjects of the GAW12 data to carry out a case-control study and a case-case study. The outcome variable chosen was the disease occurrence indicated by the dichotomous liability in the data set. The independent variables used were gene 1, age at examination, gender, the environmental factors (E1 and E2), and household effect (HH).

Case-Control Approach

Cases and controls were randomly selected from the population isolate. The sampling procedure had two steps. In the first step, two random samples of replicates were selected from which the cases and the controls would be selected respectively in the next step. Each random sample contained 15 replicates (Table I) (sampling fraction = 0.3). In the second step, within each replicate, about 10% of those with the disease were selected as our cases; and about 5% of those without the disease were selected as our controls resulting in 435 cases and 539 controls selected for the study. The selection within each replicate was independent of the size of the pedigree. The overall sampling probability was equal for all the subjects since each replicate had equal numbers of subjects. Sixteen percent of cases and 13% of controls were first-degree relatives. Table I shows the distribution of related cases and controls by degree of relationship. The distribution of the related subjects among the cases and the controls was similar.

The genetic information for the site "557" on gene 1 was extracted from the sequence data. Subjects (12%) with no copies of the mutant allele at this site were classified as the reference group (gene 1 = 0). Subjects with one or two copies of the mutant allele at this site were classified as the exposed group (gene 1 = 1). Since

TABLE I. Distributions of Selected Variables Among Cases and Controls

Variable	Case	Control
Number	435	539
Replicates selected	5, 9, 10, 11, 16, 17, 22, 23 32, 33, 34, 36, 44, 47, 48	5, 6, 9, 10, 11, 22, 23, 26 32, 34, 37, 38, 40, 41, 49
Male:Female	131:304	300:239
Degree of kinship		
first	69	70
second	42	49
third	69	63
Mean age at exam	50.25	43.18
Subjects with ancestral gene 1	97	21
E1 (top 25%)	143	101
E2 (1)	173	214

there were nearly equal numbers of subjects who carried one (444) or two (412) copies of the mutant (nonancestral) allele, additional dummy variables were created in the analysis to study the disease occurrence among those who carried different numbers of alleles.

The environmental factor (E1) was dichotomized and the cut-off point was set at the upper 25% since the value for E1, which measures the exposure, increases sharply after this point. Subjects with a value of E1 in the upper 25% were, therefore, coded as exposed.

Case-Case Approach

We further examined the interaction between gene 1 and sex in a case-case approach. This analysis focused on the cases only (435 subjects). With this analysis, main effects of the gene 1 and sex cannot be estimated, but a measure of association between the two factors among cases is easily computed as a cross-product ratio [Piegorsch et al., 1994; Khoury and Flanders, 1996; Schmidt and Schaid, 1999]. The cross-product ratio depends on both the interaction and population association of gene 1 and sex. If these two factors are independent in the population, the cross-product ratio is equal to a risk ratio. If, in addition, the disease risk is small at all levels of sex and gene 1, the ratio is approximately equal to the odds ratio for gene×environment interaction that is computed from case-control data [Schmidt and Schaid, 1999] (e.g., logit p(x=sex) = $\alpha + \beta_1$*gene 1). The independence assumption for gene 1 and sex was tested in the control subjects, using a χ^2 test.

Statistical Analysis

The association between the disease liability and the genetic factor was examined while the possible effects from other factors such as sex, environmental factors (E1 and E2), and age at examination were controlled. We used a logistic regression model [SAS Inc., 1997] to estimate the odds ratios (ORs) measuring both main effects and interaction (models 1-3). The interaction was defined as the joint effect on the disease liability of two factors, in this study, gene 1 and sex (gene 1×sex). The interaction between gene 1 and sex was studied on a multiplicative scale in both the case-control and case-case analyses.

RESULTS

A total of 435 cases and 539 controls were selected from the population through the sampling procedure described above. The sex ratios in the cases and controls from the study sample were similar to that in the population isolate. Although they were correlated in the population, there was no correlation between the household and the environmental factor (E1) among our randomly selected subjects [OR = 1.000 (0.999, 1.001)]. These findings suggested that our randomization was successful. Table I summarizes the distributions of the variables in cases and controls.

We first examined the crude effects of all the independent variables on the disease liability. In the crude analysis, each independent variable was put in the model one at a time. The crude ORs of all explanatory variables are given in the first column of Table II. Except for the variables E2 and household, every variable was significant in

TABLE II. Association Between Disease and Genetic and Environmental Factors in Crude and Adjusted Case-Control Analyses

Variable	OR_{Crude} (95% CI)[a]	OR_{Adj} (95% CI)[b]
Gene 1[c]	0.252 (0.201, 0.316)	0.216 (0.168, 0.278)
E1	2.124 (1.582, 2.852)	2.244 (1.591, 3.165)
E2	1.003 (0.775, 1.298)	0.889 (0.654, 1.210)
Sex	2.913 (2.232, 3.809)	3.519 (2.575, 4.808)
Age	1.027 (1.018, 1.035)	1.033 (1.023, 1.043)
Household	1.000 (0.999, 1.001)	1.001 (0.999, 1.002)

[a] CI, confidence interval.

[b] Each variable was adjusted for the effects of the other listed variables. Model 1: logit p(x = disease) = $\alpha + \beta_1$*sex+ β_2*gene 1 + β_3*age + β_4*E1 + β_5*E2 + β_6*household.

[c] Gene 1 coded as 0, 1, and 2.

predicting the disease when all other covariates were ignored. To control for the confounding effects among the variables, we carried out an adjusted analysis in which the effects of the genetic factor (gene 1), the environmental factors (E1, E2), sex, age at examination, and household were examined simultaneously. The column OR_{adj} in Table II gives the summarized ORs. For example, the adjusted OR for sex measured the effect of sex while all other factors were controlled (model 1 in Table II). All variables except E2 and household were significant (column 2 of Table II). Finally, we examined gene×environment interaction (gene 1×sex) while E1 and age at examination were controlled (i.e., adjusted for in the analysis) (Table III, model 2). The associations between the disease and gene 1, E1, age at examination, and sex were significant. However, we did not find significant evidence for interaction on the multiplicative scale between gene 1 and sex.

Before we carried out the case-case analysis, we examined whether gene 1 and sex were correlated among the control subjects. Gene 1 and sex were not associated in the controls (χ^2 = 2.267, p = 0.322) and therefore the assumption of independence between the genetic factor (gene 1) and sex in the case-case approach was valid. In our case-case analysis, we chose sex to be the response variable in the logistic model (Table IV, model 3). We found a similar but significant result (OR_{int} = 1.415, 95% CI: 1.030, 1.944) compared with the one in the case-control study.

In our sample, there were 362 related subjects from 138 families (first, second, and third degree relatives). To see whether correlation among subjects would affect the outcome, we excluded these subjects and reanalyzed the data using only the 612 unrelated subjects. The ORs were similar to those seen in Tables II to IV (data not shown).

TABLE III. Disease Odds Ratios and 95% CI for Gene×Environment Interaction (Gene 1×Sex) in the Case-Control Design

Variable	OR_{Adj} (95% CI)[a]
Gene 1	0.189 (0.127, 0.282)
E1	2.225 (1.578, 3.136)
Sex	2.624 (1.304, 5.277)
Age	1.032 (1.022, 1.042)
Gene 1×sex	1.254 (0.755, 2.082)

[a] Model 2: logit p(x = disease) = $\alpha + \beta_1$*sex+ β_2*gene 1 + β_3*age + β_4*E1 + β_5*gene 1*sex.

TABLE IV. Odds Ratios and 95% CI for Gene×Environment Interaction (Gene 1×Sex) in the Case-Case Design

Variable	OR$_{Adj}$ (95% CI)[a]
Gene 1	1.415 (1.030, 1.944)
E1	0.865 (0.558, 1.341)
Age	1.008 (0.995, 1.021)

[a] Model 3: logit p(x = disease) = $\alpha + \beta_1$*gene 1 + β_2*age + β_3*E1.

DISCUSSION

We conducted a population-based case-control study in which the study subjects were randomly selected from the population isolate. For the GAW12 data set, this approach is a reasonable choice since the disorder under study has a relatively high frequency in the population and the applications of the traditional epidemiologic design and analytic method such as a case-control design and the logistic model are valid [Kleinbaum et al., 1982]. We further carried out a case-case analysis using the cases that were selected in the case-control approach. The purpose of the case-case approach was to see if there was significant and consistent evidence of interaction between the gene (gene 1) and exposure (sex).

In the case-control analysis, we found that gene 1, the environmental factor (E1), and sex were strong risk factors. The age at examination was also predictive but the effect was minor. There was no evidence for gene 1×sex interaction and this could be due to the sample size of this study. These findings were consistent with the simulated data. We chose to examine interaction between gene and sex because such an interaction had been observed in a preliminary combined segregation-linkage analysis using the GAW12 data [Martinez et al., 2001]

To increase the power of detecting gene×environment interaction, we carried out a case-case analysis. We found a significant multiplicative interaction effect between gene 1 and sex (Table IV). In order to see whether this finding was due to random error of sampling, we used the same selection procedure to take another random sample of cases from the population isolate. Using the case-case approach on this second sample showed similar results [OR$_{int}$ = 1.645 (1.181, 2.292)].

Since our analytic approach required examination of multiplicative interaction, the explicit simulation of no interaction between gene 1 and sex on an additive scale had major implications for our results. Because both gene 1 and sex had effects on disease liability, at most one of the ratio or difference measures of effect could be uniform across strata [Rothman and Greenland, 1998]. Therefore, in the case-case analysis, which had more power to detect interaction than the case-control analysis, significant evidence for multiplicative interaction was observed. This was to be expected because of the effects of gene 1 and sex on disease liability and the explicit uniformity of gene 1 across sex on an additive scale [Rothman and Greenland, 1998]. The issue remains as to whether the statistical interaction detected could be biologically meaningful. There is still much debate in the epidemiology literature on appropriate definitions of interaction and interpretation of interaction or effect measure modification [Khoury and Flanders, 1996; Rothman and Greenland, 1998]. Regardless of the biological meaningfulness of the interaction detected in the current study, the importance of defining the measure of gene×environment interaction is clear.

Given that in studies using real data, one does not know the true main effects or interactions a priori, results from this simulation study provide a cautionary example for interpreting results from studies in true settings where interaction could be both statistical and biological. Our epidemiologic approach permitted only examination of multiplicative interaction. We identified such interaction between gene 1 and sex because of the explicit simulation of no additive effect between these variables and because both variables had effects on disease liability. Thus, when conducting statistical analyses, one should consider biological plausibility as part of the interpretation of results [Weed and Hursting, 1998].

In conclusion, use of a population-based case-control design was able to detect a major gene 1, sex and age as predictors for the disease liability. No significant evidence for a gene×sex interaction on a multiplicative scale was detected. However, the case-case approach, which has greater power for detecting interaction on a multiplicative scale [Yang et al., 1997] detected the joint effect of the gene and sex. Both approaches were generally effective and not too computationally intensive. Thus, traditional epidemiologic approaches may be helpful in conjunction with population/molecular genetic methods in the analysis of complex diseases with both genetic and environmental components. Finally, the necessity of explicitly defining interaction and considering biologic plausibility was clearly illustrated.

REFERENCES

Khoury MJ, Beaty TH. 1994. Application of the case-control method in genetic epidemiology. Epidemiol Rev 16:134-50.

Khoury MJ, Flanders WD. 1996. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls! Am J Epidemiol 144:207-13.

Kleinbaum DG, Kupper LL, Morgenstern H. 1982. Epidemiologic research: Principles and quantitative methods. Belmont: Lifetime Learning Publications.

Martinez M, Goldstein A, O'Connell JR. 2001. Comparison of likelihood approaches for combined segregation and linkage analysis of a complex disease and a candidate gene marker under different ascertainment schemes. Genet Epidemiol, this volume.

Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153-62.

Rothman KJ, Greenland S. 1998. Modern epidemiology. Philadelphia: Lippincott-Raven Publishers.

SAS Institute, Inc. 1997. SAS/STAT software: Changes and enhancements through Release 6.12. Cary, NC: SAS Institute, Inc.

Schmidt S, Schaid DJ. 1999. Potential misinterpretation of the case-only study to assess gene-environment interaction. Am J Epidemiol 150:878-85.

Susser E, Susser M. 1989. Familial aggregation studies: A note on their epidemiologic properties. Am J Epidemiol 129:23-30.

Weed DL, Hursting SD. 1998. Biologic plausibility in causal inference: Current method and practice. Am J Epidemiol 147:415-25.

Yang Q, Khoury MJ, Flanders WD. 1997. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713-20.